Transporting Real Objects into Virtual and Augmented Environments

Catherine Taylor
University of Bath
c.taylor3@bath.ac.uk

Murray Evans
University of Bath
m.evans@bath.ac.uk

Robin McNicholas
Marshmallow Laser Feast
robin@marshmallowlaserfeast.com

Darren Cosker
University of Bath
d.p.cosker@bath.ac.uk

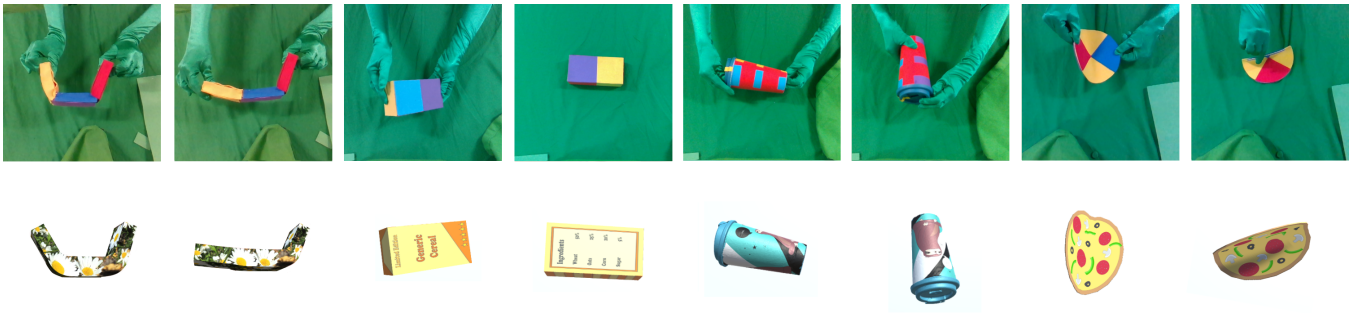

Figure 1: A physical object can be transported into a virtual environment and used as an interactive prop. These props can be used as novel controllers or as a proxy for the real object

ABSTRACT

Despite the growing interest in virtual and augmented reality (VR/AR), there are only a small number of limited approaches to transport a physical object into a virtual environment to be used within a VR or AR experience. An external sensor can be attached to an object to capture the 3D position and orientation but offers no information about the non-rigid behaviour of the object. On the other hand, sparse markers can be tracked to drive a rigged model. However, this approach is sensitive to changes in positions and occlusions and often involves costly non-standard hardware.

To address these limitations, we propose an end-to-end pipeline for creating interactive *virtual props* from real-world physical objects. Within this pipeline we explore two methods for tracking our physical objects. The first is a multi-camera RGB system which tracks the 3D centroids of the coloured parts of an object, then uses a feed-forward neural network to infer deformations from these centroids. We also propose a single RGBD camera approach using *VRProp-Net*, a custom convolutional neural network, designed for tracking rigid and non-rigid objects in unlabelled RGB images. We find both approaches to have advantages and disadvantages. While frame-rates are similar, the multi-view system offers a larger tracking volume. On the other hand, the single camera approach is more portable, does not require calibration and more accurately predicts the deformation parameters.

SCA '19, July 26 - 28, 2019, Los Angelos, CA, USA

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Modeling and simulation**; **Computer vision**.

KEYWORDS

VR Props, Non-rigid Object Tracking, Neural Networks

ACM Reference Format:
Catherine Taylor, Murray Evans, Robin McNicholas, and Darren Cosker. 2019. Transporting Real Objects into Virtual and Augmented Environments. In *Proceedings of ACM SIGGRAPH (SCA '19).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION AND RELATED WORK

Virtual and augmented reality have become prominent fields in academia, and currently generate a large amount of industrial interest across entertainment, health, engineering and education. The key requirement of a successful VR or AR experience is the feeling of immersion and the inability to easily distinguish between the real and the virtual worlds. Therefore, we design a system for tracking real-world deformable objects and showing them in virtual environments such that the user has physical props to augment their experience. We believe our approach has the potential to increase the immersion of VR and AR experiences as well as opening the door to new virtual and augmented training environments.

Hand-held controllers have been traditionally used as a tool for interacting with virtual objects, through gestures and sequences of button presses. The position and orientation of controllers, such as those used by the HTC Vive [HTC 2019] and Oculus Rift [Oculus 2019], are used along side buttons to offer increased functionality and improve immersion. Furthermore sensors (eg. the Vive Tracker) can be attached to an object to track its position and orientation. However, these trackers do not capture non-rigid object behaviour. Alternatively, Microsoft's HoloLens [Microsoft 2019] captures

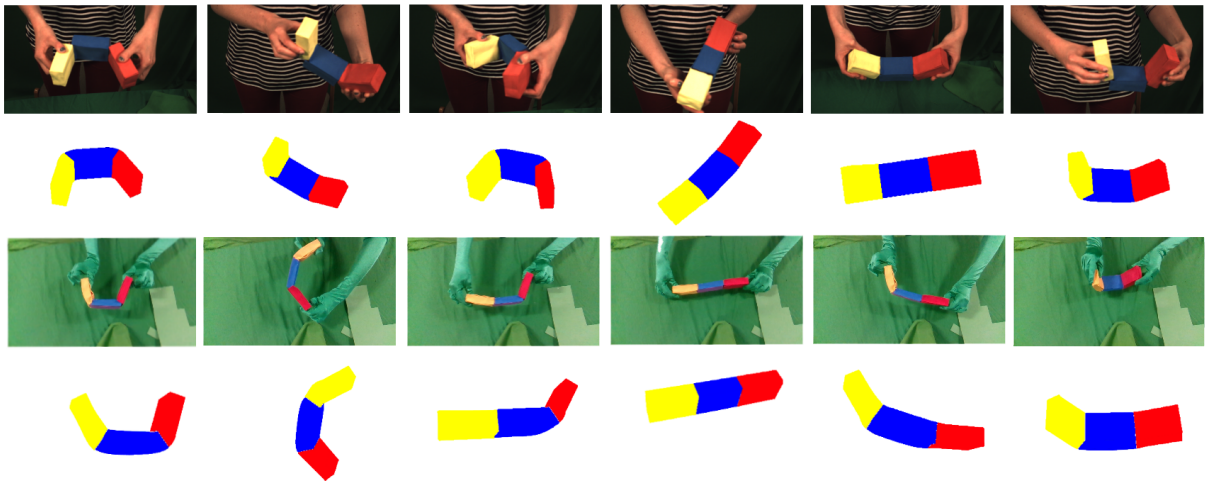

Figure 2: Predicted shape and pose on sequence of real results from both tracking approaches. The top 2 rows are from the multi-camera approach and the bottom 2 show the single camera approach.

hand gestures and uses these as means of interacting with a computer-generated object. While these methods offer increased immersion, they feel unnatural and are not the intuitive way to interact with a physical object. To capture non-rigid behaviour, markers or points on the surface of an object can be tracked and used to drive the motion of an underlying model. This is often achieved using a motion capture system, e.g. [Vicon 2019]. However, these methods are still limited as the sparseness of the markers causes the tracking to be sensitive to occlusions and fail if the positions of the markers on the object move from their initial position. Additionally, motion capture systems require specialist hardware and so can be a costly solution.

In recent years, neural networks have become an important part of computer vision and have been used within several key approaches for tracking non-rigid objects [Andrychowicz et al. 2018; Kanazawa et al. 2018; Pumarola et al. 2018]. However, these approaches require a great deal of labelled training data which is difficult and time consuming to obtain for an arbitrary object.

To overcome these limitations, we propose and compare two methods for transporting physical objects into virtual environments to be used as *virtual props*. These props can be used as alternative controllers or as a proxy for the real object. Both methods allow the tracking of both rigid and non-rigid objects. The first is a discriminative feed-forward neural network which predicts rigid and non-rigid motions from 3D object centroids observed in a multi-RGB camera setup. For the second method, we propose a new neural network - *VRProp-Net* - for tracking non-rigid objects from unlabelled RGB images and describe how this can be used within a single RGBD camera system. We show results for both approaches on several rigid and non-rigid objects.

2 OUR APPROACH

In our approach, virtual representations of arbitrary physical objects are generated without manual sculpting or rigging. Using a 3D scanner, we obtain a textured polygon mesh. This is used to simulate the non-rigid behaviour of the object using finite element analysis rather than capturing real object deformations. The many simulations are reduced using Principal Component Analysis (PCA). The PCA eigen vectors at 2 standard deviations are used to create a rigged model. To transport our physical object into the virtual world, its motion must be tracked and used to drive the motion of the rigged model. We consider two methods for capturing the change in shape and pose of a non-rigid object. An advantage of tracking an object for a virtual environment is that we can ensure a controlled capture environment with the use of green screens and object texturing (see Figure 1).

Feed-Forward Multiple RGB Camera Object Tracking. For our first tracking approach, we use a calibrated system of multiple RGB cameras. Each section of the chosen object (as seen in Figure 1) can be coloured differently and tracked using colour thresholding. For each camera, the 2D centroids of the coloured blobs are calculated and these positions triangulated to obtain a set of 3D object centroids. Using the calibration matrix, a set of synthetic images from multiple view points are rendered, with the pose and shape of the object randomly varied between each frame. The virtual object must have the same dimensions and colour as the real-world object which it represents. These synthetic images are used to train a feed-forward neural network to predict pose and deformation. The network does not directly observe the images and is instead trained by passing synthetic training images through the tracking system to get 3D centroids for known object parameters and using these to train the network. It is important that the centroids come from the tracking system, and not directly from the posed synthetic object, as prediction from the tracking system is unlikely to be the same as the exact centroids of the object parts due to the simplistic colour tracking model.

Single RGBD Camera Object Tracking: VRPRop-Net. Alternatively, we consider a novel neural network based approach that predicts pose and shape from a single RGBD camera. The physical object is captured by the RGBD camera, segmented and the colours flattened to remove variance due to non-uniform lighting. The RGB image is cropped around the centroid and the centroid back projected using the camera intrinsic matrix and the average depth to find the 3D position of the object. The cropped image is used as an input to the network and the predicted deformation parameters which are returned are used to update the virtual object's shape and pose with the resulting model rendered into a virtual scene. Our architecture, VRProp-Net, is based on a wide residual network [He et al. 2016] with the number of convolutional layers in a basic block doubled from 2 to 4 and the kernel size increased from 3 to 5. This increases the power of the blocks and allows them to better learn the deformation parameters. We train VRProp-Net on a synthetic dataset, generated by randomly varying the 3D pose and blend weights of our rigged model and rendering a 2D RGB image of each pose from a single view point. Though trained on synthetic data, VRProp-Net adapts to make predictions on real images.

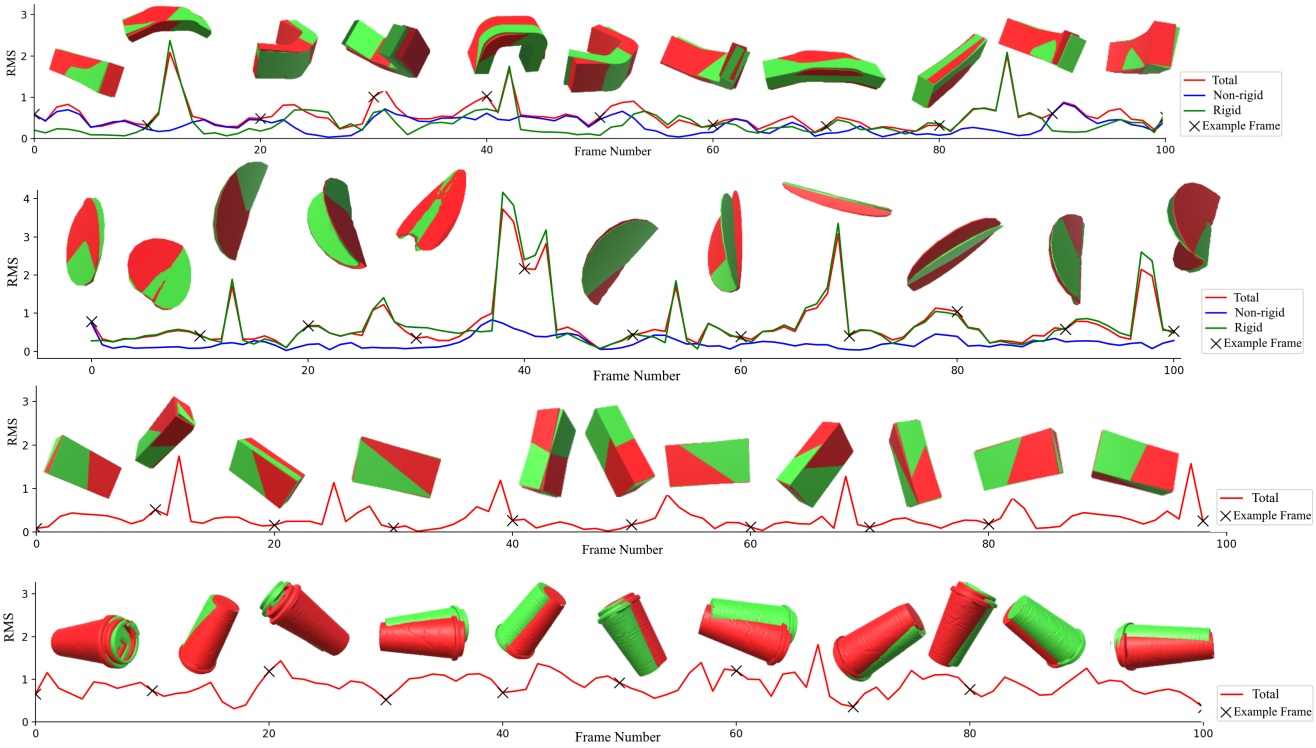

Figure 3: Predicted shape and pose on sequence of synthetic data using VRProp-Net. The RMS error between the predicted and ground truth mesh is calculated for each frame. The ground truth mesh (green) and the predicted mesh (red) are shown for a selection of frames. The total RMS error can be divided into the contributions from rigid and non-rigid transforms.

3 RESULTS

We tested our pipeline on several objects. For our multi-camera setup, we use 3 camera machine vision system which allow us to capture a volume of approximately $0.6m \times 0.6m \times 1.0m$. Using a single camera - an Intel Realsense D435 - we are able capture a smaller volume of approximately $0.9m \times 0.2m \times 0.25m$.

We compared the tracking results for each approach on a non-rigid sponge object, whose model consists of 2 blendshapes (as seen in Figure 2 and the supplementary material). Both approaches adapted well to tracking rigid and non-rigid motions on real images although trained on synthetic data. The multi-view camera system is more robust to occlusions as the additional cameras capture a wider view of the tracked object. However, VRProp-Net makes more accurate predictions. The performance of both systems is the same, with an average frame rate of around 15fps.

The ability of VRPRop-Net to predict deformation parameters from unlabelled RGB images was explored on both synthetic and real images for 2 rigid and 2 non-rigid objects. VRProp-Net was able to make accurate predictions for a range of different objects (as seen in Figures 1, 3, 4 and the supplementary material). In addition to the visual results, for the synthetic images we have the ground truth parameters and so are able calculate the Root Mean Square Error (RMS) to measure the success of the prediction. The RMS for a sequence was found to be acceptable for each object and generally have few and small changes in prediction parameters (noticeable as object jumps) between frames (see Figure 3). Figure 3 demonstrates our results on a synthetic sequence for several objects and Figures 1 and 4 shows the results on real data.

4 CONCLUSION

We have designed a pipeline to transport a physical object into a virtual environment and proposed two methods for tracking the rigid and non-rigid motions of the object. The first uses a multi-camera set up and a simple feed-forward neural network. We also propose VRProp-Net, a novel neural network for tracking rigid and non-rigid motions in unlabelled RGB images, and demonstrated its success on both synthetic and real sequences. Each approach demonstrates acceptable tracking results. However, they each have there own advantages. The multi-camera approach allows a much large tracking volume than the single camera approach and is less sensitive to occlusions. However, this system must be calibrated and the feed-forward network trained for each camera configuration as well as each object so it is a less portable solution. On the other hand, VRProp-Net is a more portable solution and makes predictions which more closely match the input image.

We have currently used our VRPRop-Net solution on a single camera. As future work, we would like to adapt the network to run on multiple cameras so that we can increase the size of the tracking volume. Additionally, we wish to explore different representations of non-rigid objects (e.g. articulated) and adapt both tracking systems to learn the parameters of these models.

REFERENCES

M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. W. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba. 2018. Learning Dexterous In-Hand Manipulation. *CoRR* (2018).

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. 630–645.

HTC. 2019. Discover Virtual Reality Beyond Imagination. https://www.vive.com/uk/.

A. Kanazawa, M. J Black, D. W Jacobs, and J. Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the CVPR*. 7122–7131.

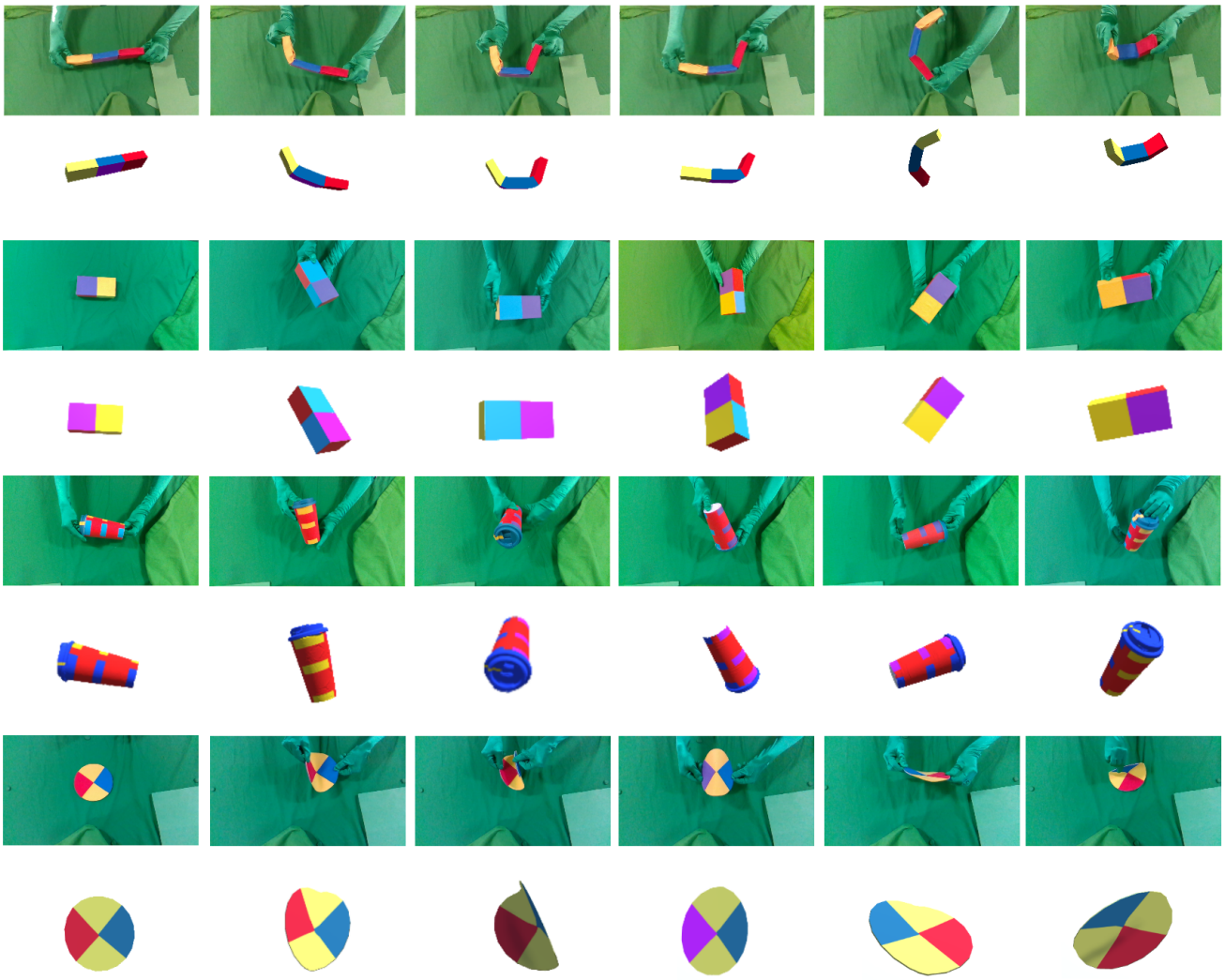

Figure 4: Predicted shape and pose on sequence of unlabelled RGB using VRProp-Net. The predicted parameters are applied to a computer generated object which is rendered into the virtual scene.

Microsoft. 2019. Microsoft HoloLens. https://www.microsoft.com/en-us/hololens.
Oculus. 2019. Oculus Rift. https://www.oculus.com/rift/.
A. Pumarola, A. Agudo, A. Porzi, L.and Sanfeliu, V. Lepetit, and F. Moreno-Noguer. 2018. Geometry-aware network for non-rigid shape prediction from a single view. In *Proceedings of CVPR*. 4681–4690.
Vicon. 2019. Motion Capture Systems. https://www.vicon.com/.